DOCUMENT RESUME

ED 346 131 TM 018 421

AUTHOR Abdel-fattah, Abdel-fattah A.
TITLE A Factor Analytic Item Response Theory Approach for Relating Item Content to Test Scores.
PUB DATE 92
NOTE 18p.; Paper presented at the Joint Annual Meeting of the Psychometric Society and the Classification Society (New Brunswick, NJ, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Difficulty Level; English; Estimation (Mathematics); *Factor Analysis; Item Analysis; *Item Response Theory; Language Tests; Matrices; Models; Probability; *Scaling; *Scores; Test Content; *Test Items; Test Results; Transformations (Mathematics)
IDENTIFIERS BILOG Computer Program; Binary Scores; LISREL Computer Program; *Ranking

ABSTRACT

A scaling procedure is proposed, based on item response theory (IRT), to fit non-hierarchical test structure as well. The binary scores of a test of English were used for calculating the probabilities of answering each item correctly. The probability matrix was factor analyzed, and the difficulty intervals or estimates corresponding to the factors were related to the total scores by scaling both the intervals (also the factor loadings) and the scores to common metrics of the test scores. The new procedure is a type of a transformation of the binary scores into more informative probabilities that can be analyzed more accurately by many statistical techniques such as factor analysis. The resulting factor loadings are used to rank order the items according to the probability of their being answered correctly. This ranking procedure was compared, in terms of meaningful item grouping, to ranking by difficulties. It is recommended that the Linear Structural Equation Relations (LISREL) computer programs be used in future studies to relate the structures of several test forms for a given administration and across administrations when supplemented with demographic variables of the students and the teachers. BILOG's Bayesian procedures for estimating item and ability parameters are also recommended because of the importance of estimation accuracy to model fit. The proposed ranking procedure should always be compared with the item difficulties ranking procedure for meaningful item grouping. One figure and four tables illustrate the analyses. There is a 15-item list of references. (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ABDEL-FATTAH A. ABDEL-FATTAH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346131

# A Factor Analytic Item Response Theory Approach for Relating Item Content to Test Scores[1,2]

Abdel-fattah A. Abdel-fattah

Educational Testing Service, 1992

[1]The opinions expressed in this chapter are those of the authors and should not be mistaken to represent the policy of the Educational Testing Services.

[2]This paper was presented for the 1991 joint meeting of the Classification Society and Psychometric Society held at Cook Campus of Rutgers University.



BEST COPY AVAILABLE

TM018421

Abstract

Many of the currently used tests are not developed to fit hierarchical, Guttman's (1954) simplex, model that relates scores of examinees to the content of the items. Unique to item response theory (IRT) is the scaling of the respondents and the items on the same continuum for hierarchical as well as nonhierarchical models. In this paper, we propose a scaling procedure based on IRT model that fits nonhierarchical test structure as well. The binary scores of a test of English were used for calculating the probabilities of answering each item correctly. The probability matrix was factor analyzed and the difficulty intervals or estimates corresponding to the factors were related to the total scores by scaling both the intervals (also the factor loadings) and the scores to common metrics of the test scores. The new procedure is a type of a transformation of the binary scores into more informative probabilities that can be analyzed more accurately by many statistical techniques such as factor analysis. The resulting factor loadings are used to rank order the items according to the probability of their being answered correctly. This ranking procedure was compared, in terms of meaningful item grouping, to the ranking by difficulties.

It is recommended that the Linear Structural Equation Relation programs (LISREL) be used in future studies to relate the structures of several test forms for a given administration and across administrations when supplemented with demographic variables of the students and teachers. BILOG's Bayesian procedures for estimating item and ability parameters are also recommended because of the importance of estimation accuracy to model fit. The proposed ranking procedure can not be expected to be always superior to the item difficulties ranking procedure, therefore; the two procedures should always be compared for meaningful item grouping. Descriptive plots for probabilities grouped by ability levels might be promising procedure for proficiency scaling.

## Introduction and Definitions

Proficiency scaling or scale anchoring is a method for defining performance along an educational scale. There are three ways of defining performance: norm-referencing, criterion-referencing, and domain-referencing (content-referencing). Norm-referencing is comparing an examinee at a given scale level to a reference group of examinees. In criterion-referencing, an external criterion measure of the same ability or a related ability is used as the reference. This reference could be the expected score or the success probability of examinees. In domain- or content-referencing, the reference is a group of test items measuring the same ability, in some meaningful sense. Thus domain- or content-referencing is linking item content to examinees ability of a given test. A test developed for norm-referencing cannot be used efficiently to reference an external criterion. Either of the norm-referenced or the criterion-referenced tests can be linked to the corresponding test scale through the appropriate method of proficiency scaling.

Beaton (1987, 1988) defined scale anchoring as the technique that allows one to describe what students whose scores at various Scholastic Aptitude Test (SAT) score levels know and can do. The National Assessment of Educational Progress (NAEP)'s definition of proficiency scaling is describing what students at selected levels know and can do. This definition adds a non-normative interpretation of the SAT while not affecting the test, its scoring, nor its scaling. In this process, different points on the SAT scale are labeled and described in terms of students' competencies or what most of the student can do at that point (Beaton, 1988). When points on scaled score metric are tied to levels of skill (or knowledge) required for examinees to achieve such score, the scale is said to be anchored. The scale produced is a single human trait or characteristic believed to account for responses of all reading tasks, math tasks, or any other given set of tasks (Mclean, 1988). It is one way to adjust for the errors of misrepresentativeness in the total test score which does not accurately reflect the student status over time or at certain point of time. Its primary purpose is to tie levels on the total score scale to significantly meaningful behavioral tasks at which a given student is proficient, (Rock, Johnson, and Pollack, 1989). Its objective is to help users and interpreters of test scores better understand what these scores mean. That is, by describing and illustrating what test takers scoring at various levels can do (Maroney, 1990). Thus proficiency scaling is thus a process of linking, tieing, or behavioral anchoring of item contents (cognitive skills or tasks) to student abilities represented by their scaled scores.

## The Need for Proficiency Scaling

To provide a global view of students mathematics achievement, NEAP reports the average proficiency on a given scale; however, this average does not reveal the specifics of what the students know and can do. NEAP also uses the results from several grades to define skills, knowledge, and understanding that characterize a number of scale levels of mathematics performance. To define these skills that characterize each proficiency level, mathematics

specialists studied the questions that were typically answered correctly by most students at a particular level but were answered incorrectly by a majority of students at the next lower level. Types of abilities needed to answer each set of questions are then summarized. The proficiency levels were determined prior to grouping items that have similar content within each interval. NAEP thus assumes that the students knowledge is stage dependent, that the stages or intervals are equally spaced, and that item grouping will fit these equally spaced levels. These assumptions cannot be true for tests with various data structures.

Siegler (1983) argued that much of the child's cognitive development is the sequential acquisition of increasingly powerful rules for solving problems. This type of learning acquisition appears to be less stage dependent than competing models such as those of Guttman (1954), Piaget (1972), Flavell (1983), and Brown and et. al. (1983). Guttman's model is the hierarchical model discussed earlier. Piaget (1972) argued that early childhood, middle childhood, and adolescence cognitive systems are qualitatively different from one another. Piaget's performance in higher stages require at least some proficiency in the lower level cognitive stages. This notion is consistent with hierarchy of levels in Guttman's scale model. Flavell (1981) and Brown et. al. (1983) argued that the Piagitian formal-operational thinking is metacognitive in nature because it involves thinking about propositions (interpropositional thinking) and requires the skill to inter-connect simpler item strategies. This type of thinking, as well as lower levels skills, is required by higher stages items.

For Siegler's model with less stage dependent acquisition, the hierarchical model does not always fit the students' cognition. Also with the equally intervalled stages we cannot sufficiently assess students knowledge whether this hierarchical model is used or not.
Exploring the structure of the data would be a better approach to investigate students cognition in terms of hierarchy or equally spaced learning stages.

Test-takers often ask about their test scores and subscores to see how well they did on the corresponding subtests and to study more for the ones they were deficient at. If every test-taker is given an idea of the subjects that need to be studied more, these test-takers will apparently improve the skills required by that test. Proficiency scales will provide students with a list of skills they need to work on based on the test score they have acquired. In a sense proficiency scale is some kind of a remedial feedback that almost every student need for skill and score improvement. This way we identify the typical content of items that respondents with $\theta$ or total score values in the neighborhood of the scale point have mastered. A number of items in the vicinity of the selected points should be examined in order to give the content interpretation some generalizability.

Test developers need to have an idea about the test structure which serves our educational goals rather than to reflect only the content of a given test (as a sample of knowledge) that may or may not reflect these goals. By figuring out the structure of the skills of a given test, test developers can change this structure or keep unchanged so that it may serve our educational goals. Psychometrician and measurement statistician can obtain tests of known structure and well-defined function so that they can prove or disprove whether these structures serve the

required goals and functions or not. They can also direct their efforts to implement theories that serve that purpose.

## The Model

The Guttman's simplex model is usually the model adopted in the process of behavioral anchoring. In this model the skills tested are assumed to have a hierarchical order, a wide range of tasks, to be discrete, and a widely represented (diverse) population. These assumptions imply that items should meet certain criteria before they are used in anchoring. Items should discriminate highly between proficiency levels and they should be homogeneous in difficulty and in content within each of these levels. This model is very restrictive and most of the available data do not fit such a model. It is one special case of the item response theory model. That is, the one-parameter logistic model with zero guessing and highly discriminating items. It is very difficult to construct tests that fit these characteristics. Therefore, it is even more difficult to use this model for tests constructed without attention to the model it may fit. These assumptions can be relaxed by using other models, such as the three-parameter logistic model. This model is the general case of which the Guttman's model is a special case. Thus hierarchical data structure is expected to fit the three-parameter logistic model but the nonhierarchical data would not fit the Guttman's simplex model without ignoring items that do not fit the model.

By using the three-parameter logistic model, we can scale the respondents and the items on the same continuum. The respondents are located on the ability continuum by their estimated scale score $\theta$; and the items are located by their threshold estimates b (which is equal to the ability point with probability of answering an item correctly equal to $(c+1)/2$, where c is the guessing parameter estimate). The difficulty which depends on the guessing parameter estimates, and the ability estimates of a given BILOG (or LOGIST) run are automatically on the same scale. Thus the only parameter estimate not included in this picture is the item discrimination parameter estimate a. Therefore, the a's, the b's, the c's, and the $\theta$'s are used in computing the probabilities of answering each item correctly by each of the examinees. Consequently, these probability contain the $\theta$'s scaled to the b's, the b's which depend on the c's, the c's, and the a's which was not previously accounted for. The probabilities can be thought of as the correlations between characteristics of a given item and the ability of each examinee to answer that item correctly. Factor analysis of these probabilities results in factor loadings that represent not only information in the difficulty, ability, and guessing parameter estimates but also information in the item discrimination estimates. Factor loadings are actually the correlations between items characteristics (a's and b's) and the overall ability continuum $\theta$.

## Summary of Previous Work

There is a small body of literature relevant to the use of hierarchical models on nonhierarchical tests. Previous work on that subject is reviewed and discussed in this section. The three procedures currently used in obtaining proficiency scales are the empirical, the model-

based, and the Livingston-modified procedure. In the empirical method the number and the width of the score levels are first specified without defining the cognitive model. The second step is the estimation of item statistics such as the probabilities of correct responses within 20 units of the target level. The third step is the grouping of items with the desired characteristic (at a given level p is at least .7, at the lower level $p < .5$, the difference is at least .3). This difference in p is believed to assure that items are highly discriminating at these levels. The final step is to describe items of each score level and to identify the problems that students can or cannot do at each level.

For the model-based method, the hierarchical cognitive model is first defined (e.g., three skill levels). Testlets or clusters of items were used, instead of single items, to anchor the scale. These testlets were selected according to how well they represent one of the cognitive level skills and how their IRT difficulty estimates are consistent with the skill. Only the four items with the highest discrimination were selected to assure that these items discriminated well at that skill. At each of the three levels a super item is defined and used in assuring model hierarchy (see page 2 of Rock et al report for how it was used). The IRT individual probabilities of proficiency at each of the three levels for each super item can be tied to individual's theta score.

Livingston (1990) suggested that three assertions should be demonstrated regarding proficiency scaling: using different test forms and different experts to describe the score levels, a particular score level does not imply a particular response pattern (skill or proficiency), and all or nearly all items are classified as measuring certain skill will discriminate well at that level. In his scaling for the proficiencies in the Russian test, Livingston found out that the use of more experts was not better because of contradictory classifications by these experts to each other's and to the item difficulties. He also found out that some discrimination estimates were so small because of the inaccurately estimated abilities. He imposed two restrictions on the data, these are: the constant guessing estimates and the constant probability of .70 for answering an item correctly within a given level. These conditions may not be true for all items in the test.

Livingston (1990) scaled the proficiencies of a Russian Language Test (RLT). He accounted for almost all the assertions that he suggested earlier. Even after he accounted for these assertions, he was met with other problems. For instance, the item ratings were contradictory to the item parameter estimates and contradictory from one rater to the other. The new problems are accounted for in the current study.

## The New Procedure

The three approaches mentioned earlier are based on the Guttman's simplex hierarchical model. To use these models, we have to maximize item discrimination within each cognitive level. This is done by choosing highly discriminating items with similar difficulties, as in the empirical method, or by using highly discriminating super items as in the model-based method. Choosing highly discriminating items does not guarantee that each interval is distinct and discriminated from the other intervals. It is the difference in discrimination from one interval

to the other that counts. Besides there is no way to get all or almost all items with high discrimination from a test like SAT, because the test was not developed that way. The arbitrary 100 score points wide interval is unrealistic because a given score can be attributed to more than one skill or subject area leading to the overlap of these intervals. The different number of items participating in each skill also leads to varying the widths of these intervals. The way we figured out to obtain these interval widths and the amount of overlap is the exploratory factor analysis of the probabilities.

The suggested approach is one in which we may account for Livingston's latest concerns in his analysis of the Russian Language Test. The reason for not using the empirical procedure is its lack of validity. The model-based approach is more valid but still restrictive to the type of items and the scale intervals being used. The new approach is less restrictive and more exploratory and believed to be an improvement over the other approaches. In the following section the new procedure will be introduced and presented.

Item content can be described in terms of the IRT item and ability parameter estimates, the classical test score item statistics, and or the ratings of test developers. The probabilities of an item i answered correctly by an examinee j are calculated from the IRT parameter estimates. The ability and the item parameter estimates should be as credible as possible otherwise the scaling would not be as accurate. The procedures believed to produce most credible estimates are the hierarchical Bayesian estimation procedure for item and ability parameter estimation. These procedures are recommended to be used in this study to estimate the item and ability parameters. Because of time constrains and cost limitations, LOGIST estimates were used instead. A matrix of probabilities is calculated from the item and ability parameter estimates. The number of rows and columns is the number of examinees and items; respectively. These probabilities can be used to tie the test scale scores (that describe examinees) to the cognitive model levels (that describe items).

The cognitive model levels (CMLs) can be obtained by classifying items into clusters based on their content, complexity, and/or item parameter estimates. The classification procedure is preferred to be numerical in nature because the experts' ratings have been proven, at least for some items, contradictory (Livingston, 1990) not only to each other's but also to the difficulty estimates. The proposed classification procedure is the factor analysis for the probability matrix. The use of these probabilities allows the inclusion of item and ability parameter estimates in the classification procedure because these probabilities are calculated from these estimates. The number and the width of the CMLs are only determined numerically as the data is analyzed not before the analysis. This way the data provide the guidelines rather than individual judgement which was proven contradictory.

The ability estimates and the difficulty estimates are scaled to the same metrics when they are produced by the same LOGIST (or BILOG) run. Either of these estimates is not on a common metric with test scores in a given form or across forms. Therefore, the difficulty estimates are transformed to the scale scores of a given test across several forms by using any

equating program (e.g., chi-square by Divgi, 1985) which equate different forms to a common scale.

## Data Collection and Analysis

The binary scores of 3000 students on an English test were used as the input for LOGIST. The test constitute two sections. The first section (50 items) was developed to test the students' ability to recognize and understand idioms by answering questions after listening to a short conversation. The second section (40 items) tests the ability of the students on English structure and vocabulary. The resulting a, b, c, and $\theta$ estimates were used for calculating the examinees by items probability matrix. The probability matrix is factor analyzed by using proc-factor SAS procedure. The resulting factors and factor loadings are used for anchoring item content to the test scores by describing items within each factor or group of items. The steps of the analysis and scaling proficiency of a test are summarized in Figure 1.

---

INSERT FIGURE 1 HERE

---

## Results, Discussion, and Conclusion

Factor analysis of the probability matrix for items 1 to 50 resulted in a two-factors solution shown in Table 1. Items 48, 1, 5, 47, 49, 21, 19, 50, 2, 17, 20, 39, 40, 16, 18, 44, 22, 4, 15, 24, 3, 28, 35, 43, and 6 loaded highly on the first factor; the rest of the items loaded highly on the second factor. The two factors did not represent two distinct groups of items that clearly differ in content. The two groups of items can be generally described to represent two levels of difficulties as confirmed by difficulties scaled to 0-100 scale as shown in Table 1. Items of the first group are those items loaded highly on the first factor; items of the second group are those items loaded highly on the second factor. In a few cases, difficulties did not agree with loadings. For example, item 39 had a negative difficulty (Table 2) but loaded highly on the first factor (Table 1).

This part of the test was developed considering that the items should represent a problem in understanding English expressions. Distinctive hierarchy in item difficulties was not taken into consideration. Therefore, item loadings may overlap the same way test scores may. For instance, items 13 and item 47 have the same scaled loading of 69 but on two different factors. We cannot say which of the two items has participated, together with items lower in rank, in the score of 69.

INSERT TABLES 1 & 2 HERE

In spite of such overlap we can generally say that students with high scores on that section of the test are expected to answer items with high loadings on either factors. The subscore corresponding to each factor should be reported, if we want to differentiate which items on either factors resulted in the overlapping score.

Factor analysis of the probability matrix for items 51 to 90 resulted in a two factors solution as shown in Table 3. Items 85, 65, 79, 90, 82, 83, 69, 60, 59, 63, 80, 67, 52, 53, 51, and 73 have high loadings on the second factor; the rest of the items loaded highly on the first factor. The two factors did not represent two distinct groups of items that clearly differ in content. Table 4 suggests that the two factors correspond to two difficulty levels as was the case in the first section of the test. Difficulties tend to have a similar rank to the factor loading. The scaled difficulties correspond to total scores because they are placed on a common metric. Thus an examinee with a score of 75 (scaled difficulty for item 74) can answer item 74 and all items easier than this item within the first factor. An examinee with score of 67 (scaled difficulty for item 58) can answer this item and all easier items. This examinee need to work on the items more difficult than item 58 within the first factor. The problem with difficulty ranks is that they ignore the possibility that two items with two different skills or content may be equal in difficulty and produce the same score. In a few cases, difficulties did not agree with loadings. For example, item 65 had a negative difficulty (Table 3) but loaded highly on the second factor (Table 4).

This section of the test was developed considering only that items should represent a problem in English structure disregarding the imposition of hierarchical skill structure into the items. Therefore, item loadings may overlap the same way test scores may. For instance, items 61 and item 82 have the same scaled loading of 92 but the loadings were on two different factors. We can only say that both items has participated in the score of 92. In spite of such overlap we can generally say that students with high scores on that section of the test will most probably answer items with high loadings on either factors. The subscores corresponding to each factor (group of items) should be reported, if we want to identify items that resulted in the same score.

INSERT TABLES 3 & 4 HERE

The data structure was obviously nonhierarchical in the sense that each item is not necessarily more complicated by one task than another. In spite such nonhierarchical data more

than one factor emerged. It is also interesting that each of the two sections are known for their unidimensionality when the binary scores were factor analyzed. The emergence of a second factor in the two sections makes one wonder about the dimensionality and the actual structure of the sections of this test.

The unidimensionality problem should not concern IRT pros since the binary scores (not the probability of these scores) were proven unidimensional which is assumed by the theory. Besides, multidimensional tests can be analyzed by IRT programs that does not assume unidimensionality. The structure of the test sections was obviously explained in a way different, more detailed, and more realistic than in previous studies. However; results from old procedures and current procedure need to be compared and possibly used simultaneously to better understand the test structure.

The use of forms parallel to the one investigated in this study may be useful in the generalization of the scaling of this form to other forms. However, because of time constrains only one form was used to demonstrate, propose, and discuss the use of the new methodology for scaling proficiency.

In future studies, BILOG's Bayes estimates are recommended instead of LOGIST's because credibility of these estimates is essential for the analysis. In this paper, the factor analysis for probabilities of answering items correctly could be criticized because of two problems. The first is that we used a linear factor analysis procedure to analyze non-linear data. The solution to this problem is to use a non-linear factor analysis model which is very expensive and does not provide a solution for the second problem mentioned below. The second is that the probabilities of answering items correctly will produce one or more difficulty factors which are indistinguishable from the proficiencies being explored. This problem is not unique to probabilities of answering items correctly but also (even more so) to dichotomous item responses. My solution and future recommendation to this problem is the use of descriptive plots for the probabilities of answering items correctly. The probabilities may be plotted against and grouped by ability levels. These plots are expected to reveal item clusters and at the same time relate these clusters to the ability levels without the complications and requirement of linearity of data nor the problem of high loadings on difficulty factors.

Figure 1. Flow chart for scaling the proficiency of a test using the factor analytic item response theory approach.

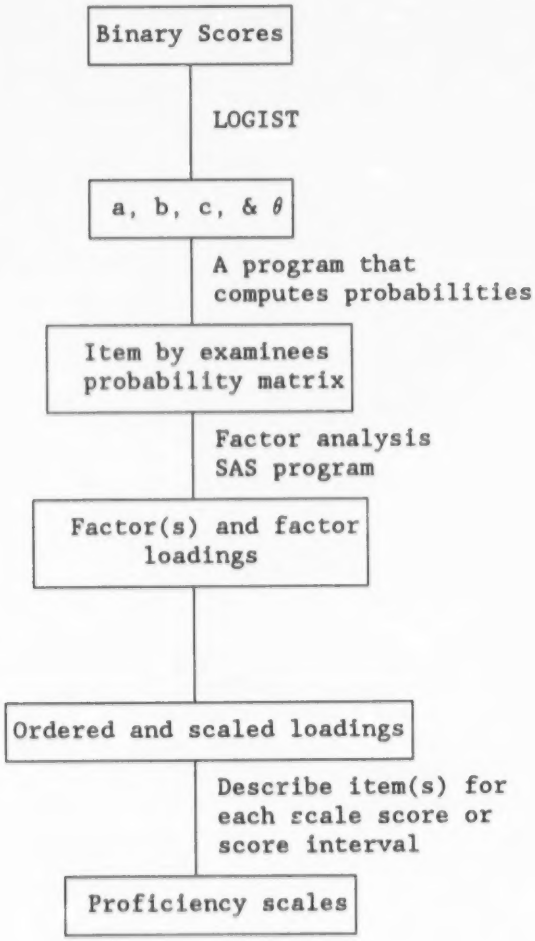
Binary Scores
LOGIST
a, b, c, & θ
A program that computes probabilities
Item by examinees probability matrix
Factor analysis SAS program
Factor(s) and factor loadings
Ordered and scaled loadings
Describe item(s) for each scale score or score interval
Proficiency scales

Table 1. Factor Loadings for Items 1 to 50 of the English Language Test.

| Items | Factor1 | Factor2 | RSB |
|---|---|---|---|
| I48 | 109 * | -1 | 29 |
| I1 | 107 * | 1 | 21 |
| I5 | 97 * | 6 | 29 |
| I47 | 96 * | 7 | 14 |
| I49 | 92 * | 12 | 27 |
| I21 | 92 * | 11 | 15 |
| I19 | 88 * | 12 | 7 |
| I50 | 85 * | 16 | 30 |
| I2 | 81 * | 19 | 9 |
| I17 | 78 * | 26 | 26 |
| I20 | 77 * | 26 | 20 |
| I39 | 75 * | 21 | .20 |
| I40 | 75 * | 24 | 33 |
| I16 | 72 * | 28 | 33 |
| I18 | 71 * | 29 | 11 |
| I44 | 62 * | 39 * | 36 |
| I22 | 61 * | 39 * | 37 |
| I4 | 61 * | 39 * | 38 |
| I15 | 59 * | 41 * | 18 |
| I24 | 57 * | 45 * | 26 |
| I3 | 57 * | 36 | 41 |
| I28 | 53 * | 46 * | 41 |
| I35 | 50 * | 42 * | 44 |
| I43 | 48 * | 38 * | 44 |
| I6 | 46 * | 35 | 45 |
| I23 | -1 | 113 * | 74 |
| I12 | 4 | 101 * | 62 |
| I13 | 0 | 96 * | 81 |
| I46 | 11 | 91 * | 70 |
| I33 | 16 | 85 * | 62 |
| I38 | 13 | 82 * | 73 |
| I14 | 15 | 79 * | 59 |
| I45 | 23 | 78 * | 59 |
| I8 | 25 | 73 * | 57 |
| I37 | 27 | 72 * | 56 |
| I25 | 27 | 70 * | 55 |
| I29 | 26 | 68 * | 55 |
| I10 | 37 * | 65 * | 50 |
| I30 | 36 | 64 * | 51 |
| I31 | 18 | 63 * | 55 |
| I26 | 42 * | 57 * | 47 |
| I42 | 32 | 56 * | 51 |
| I36 | 21 | 54 * | 53 |
| I11 | 48 * | 53 * | 43 |
| I7 | 31 | 53 * | 51 |
| I9 | 45 * | 52 * | 28 |
| I41 | 41 * | 52 * | 48 |
| I34 | 41 * | 51 * | 47 |
| I32 | 29 | 50 * | 51 |
| I27 | 37 * | 48 * | 56 |

All factor loadings are multiplied by 100 and rounded to the nearest integer.
Values greater than 36 have been flagged by an '*'.
RSB: rounded scaled difficulties.

Table 2. Items 1 to 50 Ranked by difficulty Estimates.

| | OBS | N | SB | B | A | C |
|---|---|---|---|---|---|---|
| First factor | | | | | | |
| | 1 | 39 | 0.2054 | -2.2882 | 0.47893 | 0.182673 |
| | 2 | 19 | 6.6895 | -2.0156 | 0.62728 | 0.182673 |
| | 3 | 2 | 9.0681 | -1.9173 | 0.58578 | 0.182673 |
| | 4 | 18 | 11.1251 | -1.8350 | 0.50654 | 0.182673 |
| | 5 | 47 | 13.9520 | -1.7219 | 0.79106 | 0.182673 |
| | 6 | 21 | 14.8244 | -1.6870 | 0.74918 | 0.182673 |
| | 7 | 15 | 17.8021 | -1.5679 | 0.44247 | 0.182673 |
| | 8 | 20 | 20.2744 | -1.4690 | 0.65107 | 0.182673 |
| | 9 | 1 | 21.0984 | -1.4361 | 1.11800 | 0.388671 |
| | 10 | 24 | 26.3899 | -1.2244 | 0.49289 | 0.182673 |
| | 11 | 17 | 26.3906 | -1.2244 | 0.77686 | 0.182673 |
| | 12 | 49 | 26.8878 | -1.2045 | 1.06457 | 0.111303 |
| | 13 | 9 | 27.8867 | -1.1645 | 0.37497 | 0.182673 |
| | 14 | 48 | 28.6753 | -1.1330 | 1.70000 | 0.269288 |
| | 15 | 5 | 29.4673 | -1.1013 | 1.35675 | 0.296819 |
| | 16 | 50 | 29.7665 | -1.0893 | 1.07129 | 0.049753 |
| | 17 | 16 | 33.1124 | -0.9555 | 0.91854 | 0.360368 |
| | 18 | 40 | 33.3773 | -0.9449 | 1.02200 | 0.115004 |
| | 19 | 44 | 36.0782 | -0.8369 | 0.79308 | 0.077286 |
| | 20 | 22 | 37.0236 | -0.7991 | 0.83521 | 0.234185 |
| | 21 | 4 | 37.5425 | -0.7783 | 0.84879 | 0.111675 |
| | 22 | 3 | 40.7384 | -0.6505 | 1.07277 | 0.207673 |
| | 23 | 28 | 40.9444 | -0.6422 | 0.81994 | 0.044915 |
| | 24 | 11 | 43.4572 | -0.5417 | 0.73783 | 0.023106 |
| | 25 | 35 | 43.6480 | -0.5341 | 1.07767 | 0.119066 |
| | 26 | 43 | 44.2227 | -0.5111 | 1.23983 | 0.099691 |
| | 27 | 6 | 44.8546 | -0.4858 | 1.38145 | 0.435435 |
| Second factor | | | | | | |
| | 28 | 34 | 47.3014 | -0.3879 | 1.04624 | 0.164156 |
| | 29 | 26 | 47.4690 | -0.3812 | 0.84349 | 0.381901 |
| | 30 | 41 | 47.5184 | -0.3793 | 1.03922 | 0.154739 |
| | 31 | 27 | 48.2528 | -0.3499 | 1.24447 | 0.230694 |
| | 32 | 10 | 49.9496 | -0.2820 | 0.61102 | 0.000491 |
| | 33 | 32 | 50.4984 | -0.2601 | 1.43657 | 0.153113 |
| | 34 | 7 | 50.7040 | -0.2518 | 1.31390 | 0.156173 |
| | 35 | 42 | 50.9062 | -0.2438 | 1.17320 | 0.225332 |
| | 36 | 30 | 51.0599 | -0.2376 | 0.80893 | 0.096960 |
| | 37 | 36 | 52.8623 | -0.1655 | 1.60629 | 0.171646 |
| | 38 | 29 | 54.9305 | -0.0828 | 1.07290 | 0.204150 |
| | 39 | 31 | 55.0199 | -0.0792 | 1.49568 | 0.292665 |
| | 40 | 25 | 55.1810 | -0.0736 | 0.99169 | 0.389889 |
| | 41 | 37 | 55.6927 | -0.0523 | 0.90286 | 0.160672 |
| | 42 | 8 | 56.6561 | -0.0138 | 0.95859 | 0.393533 |
| | 43 | 14 | 59.1949 | 0.0878 | 1.27729 | 0.153783 |
| | 44 | 45 | 59.4574 | 0.0983 | 0.80794 | 0.218454 |
| | 45 | 33 | 61.8821 | 0.1953 | 1.00077 | 0.270160 |
| | 46 | 12 | 61.9831 | 0.3993 | 1.32648 | 0.260797 |
| | 47 | 46 | 69.5448 | 0.5018 | 0.75171 | 0.084061 |
| | 48 | 38 | 72.8687 | 0.6347 | 0.49532 | 0.046311 |
| | 49 | 23 | 73.5098 | 0.6604 | 1.23002 | 0.529266 |
| | 50 | 13 | 81.3667 | 0.9747 | 0.65040 | 0.345565 |

OBS: observation code
A: discrimination estimates
B: difficulty estimates
BS: difficulties scaled to 0-100 scale
C: guessing estimates
N; item number

Table 3. Factor Loadings for Items 51 to 90 of the English Language Test.

| Item | Factor1 | Factor2 | RSB |
|---|---|---|---|
| I74 | 126 * | 1 | 74 |
| I62 | 110 * | 3 | 70 |
| I89 | 104 * | 6 | 70 |
| I70 | 100 * | 6 | 67 |
| I58 | 97 * | -7 | 94 |
| I87 | 96 * | 1 | 88 |
| I57 | 95 * | 11 | 71 |
| I56 | 93 * | 5 | 63 |
| I61 | 92 * | 12 | 71 |
| I72 | 91 * | 13 | 69 |
| I86 | 89 * | 3 | 61 |
| I76 | 80 * | 22 | 63 |
| I71 | 79 * | 21 | 61 |
| I68 | 70 * | 27 | 56 |
| I64 | 67 * | 9 | 55 |
| I66 | 67 * | 35 | 54 |
| I88 | 66 * | 32 | 53 |
| I54 | 66 * | 29 | 53 |
| I81 | 59 * | 22 | 52 |
| I77 | 57 * | 28 | 51 |
| I84 | 57 * | 39 | 49 |
| I75 | 54 * | 30 | 50 |
| I55 | 49 * | 43 * | 46 |
| I78 | 47 * | 44 * | 45 |
| I85 | -5 | 116 * | 13 |
| I65 | -1 | 113 * | .54 |
| I79 | 0 | 106 * | 13 |
| I90 | 0 | 106 * | 17 |
| I82 | 11 | 92 * | 20 |
| I83 | 14 | 89 * | 18 |
| I69 | 13 | 84 * | 29 |
| I60 | 3 | 74 * | 36 |
| I59 | 37 | 70 * | 14 |
| I63 | 16 | 67 * | 37 |
| I80 | 30 | 67 * | 33 |
| I67 | 37 | 66 * | 29 |
| I52 | 31 | 63 * | 36 |
| I53 | 45 * | 56 * | 36 |
| I51 | 27 | 55 * | 41 |
| I73 | 48 * | 50 * | 42 |

All factor loadings are multiplied by 100 and rounded to the nearest integer.
Values greater than 36 have been flagged by an '*'.
RSB: rounded scaled difficulties.

Table 4. Items 51 to 90 Ranked by Difficulty Estimates.

| | OBS | N | SB | B | A | C |
|---|---|---|---|---|---|---|
| Second factor | | | | | | |
| | 1 | 65 | 0.5398 | -2.2216 | 1.24037 | 0.182673 |
| | 2 | 79 | 12.5225 | -1.6991 | 1.04738 | 0.182673 |
| | 3 | 85 | 13.1757 | -1.6730 | 1.29884 | 0.160928 |
| | 4 | 59 | 13.5521 | -1.6579 | 0.51413 | 0.182673 |
| | 5 | 90 | 17.0230 | -1.5191 | 1.13108 | 0.161480 |
| | 6 | 83 | 18.9399 | -1.4424 | 0.86116 | 0.182673 |
| | 7 | 82 | 19.8991 | -1.4040 | 0.93155 | 0.056917 |
| | 8 | 69 | 29.1118 | -1.1155 | 1.03284 | 0.144316 |
| | 9 | 67 | 29.1889 | -1.1124 | 0.65490 | 0.182673 |
| | 10 | 80 | 33.0265 | -0.8789 | 0.87867 | 0.310773 |
| | 11 | 52 | 35.7292 | -0.7708 | 0.92239 | 0.266017 |
| | 12 | 53 | 36.1384 | -0.7545 | 0.67012 | 0.077825 |
| | 13 | 60 | 36.2915 | -0.7483 | 1.62657 | 0.345685 |
| | 14 | 63 | 36.8875 | -0.7245 | 1.30585 | 0.275737 |
| | 15 | 51 | 40.6881 | -0.5725 | 1.24181 | 0.283957 |
| | 16 | 73 | 41.8534 | -0.5259 | 0.77641 | 0.167856 |
| First factor | | | | | | |
| | 17 | 78 | 45.1853 | -0.3926 | 0.98977 | 0.155754 |
| | 18 | 55 | 45.5583 | -0.3777 | 0.92923 | 0.058054 |
| | 19 | 84 | 48.6076 | -0.2557 | 0.84034 | 0.332980 |
| | 20 | 75 | 50.2766 | -0.1889 | 1.18828 | 0.295029 |
| | 21 | 77 | 51.3947 | -0.1442 | 1.19576 | 0.267331 |
| | 22 | 81 | 52.7078 | -0.0917 | 1.34763 | 0.201114 |
| | 23 | 88 | 53.3027 | -0.0679 | 0.81518 | 0.161947 |
| | 24 | 54 | 53.4837 | -0.0607 | 0.94938 | 0.229657 |
| | 25 | 66 | 54.3070 | -0.0277 | 0.51167 | 0.066106 |
| | 26 | 64 | 55.3972 | 0.0159 | 1.70000 | 0.184214 |
| | 27 | 68 | 55.6871 | 0.0275 | 0.91211 | 0.492913 |
| | 28 | 86 | 61.0798 | 0.2432 | 1.60200 | 0.097085 |
| | 29 | 71 | 61.0859 | 0.2434 | 0.85222 | 0.312984 |
| | 30 | 56 | 62.7844 | 0.3114 | 1.42374 | 0.278187 |
| | 31 | 76 | 63.0013 | 0.3201 | 0.74836 | 0.039161 |
| | 32 | 70 | 67.2472 | 0.4899 | 1.20821 | 0.222885 |
| | 33 | 72 | 68.6400 | 0.5456 | 0.87473 | 0.214181 |
| | 34 | 89 | 70.0355 | 0.6014 | 1.16693 | 0.315889 |
| | 35 | 62 | 70.2947 | 0.6118 | 1.34447 | 0.206433 |
| | 36 | 57 | 70.8622 | 0.6345 | 0.89488 | 0.246530 |
| | 37 | 61 | 71.0993 | 0.6440 | 0.82801 | 0.077246 |
| | 38 | 74 | 75.2417 | 0.8097 | 1.70000 | 0.275740 |
| | 39 | 87 | 87.9626 | 1.3185 | 0.79529 | 0.284708 |
| | 40 | 58 | 94.0437 | 1.5617 | 1.70000 | 0.375016 |

OBS: observation code
A: discrimination estimates
B: difficulty estimates
BS: difficulties scaled to 0-100 scale
C: guessing estimates
N: item number

## Bibliography


Beaton, A. E. (1988). Draft of precis for anchoring the SAT scales. An internal ETS precis draft.

Beaton, A. E. (1987). The design and analysis of NAEP data. Paper presented at an invitational seminar sponsored by the National Foundation for Educational Research and the London University Institute of Education.

Brown, A. L.; Bransford, J. D., Ferrara, R. A., and Campione, J. C. (1983). Learning, remembering and understanding. In J. H. Flavell and E. M. Markman, eds., Handbook of Child Psychology: Cognitive Development. Vol 3. NY: Wiley.

Conlan, G. (1990) Proficiency Scaling of SAT-V. An unpublished draft of an internal ETS report.

Flavell, J.H. (1983) Cognitive development during the postinfancy years. In Issues in Cognition: Proceedings of a Joint Conference in Psychology, Racine WI. National Academy of Sciences.

Guttman, L. (1954) A new approach to factor analysis: the radex. In P.F. Lazarsfeld (ed) Mathematical thinking in the social sciences. Glencoe, IL: Free Press.

Livingston, S. (1990). Review of proficiency scaling paper (by Maroney, 1990). An internal ETS memo.

Livingston, S. (1990). Scaling proficiency of Russian Language Test. An unpublished internal ETS report.

Maroney, J. M. (1990). Proficiency scaling of SAT-M. An internal ETS report.

Mclean, L. D. & Goldstein, H. (1988) The U.S. National Assessments In Reading: Reading Too Much Into Findings. Phi Delta Kappan.

Piaget, J. (1972) Intellectual evolution from adolescence to adulthood. Human Development, 15, 1-12.

Ramsay, J. O. (1988). Monotone regression splines in action (with discussion.) Statisitcal Science, 3, 425-461

Rock, D. A.; Johnson, E.; and Pollack, J. (1989) SAT mathematics proficiency scaling. An ETS report reviewed and accepted by the JRSDC as an internal report.

Rock, D. A. and Pollack, J. (1987). Measuring Gains -- A New Look at an Old Problem. Paper presented at the ETS/DOD Conference in San Diego, DA.

Siegler, R. S. (1983) Information processing approaches to cognitive development. In W. Kessen. ed., Handbook of Child Psychology: History, Theory, and Methods, Vol 1. NY: Wiley.